USE OF INTRONIC RNA TO MEASURE GENE EXPRESSION

Background of the Invention

[0001] The present invention claims the benefit under 35 U.S.C. §119(e) of provisional application Serial No. 60/448,991 filed on February 20, 2003, the entire disclosure of which is hereby expressly incorporated by reference.

Field of the Invention

[0002] It is well recognized that gene expression within cells and tissues can indicate the physiologic and or pathologic status of the cell, tissue or patient. For several decades, gene expression, as measured by immunohistochemical analysis of protein markers, has been used to make treatment decisions. For example, levels of estrogen receptor and progesterone receptor measured this way are now routinely used to select breast cancer patients for treatment with anti-estrogen drugs.

[0003] More recent research literature provides evidence that tissue levels of mRNA species have diagnostic and prognostic value. This is a promising development because the technologies for measurement of cellular RNA levels, as exemplified by multiple RT-PCR and DNA array platforms, can be very sensitive, specific and quantitative. RT-PCR is recognized as generally more sensitive than DNA array technology. However, RT-PCR probe/primer design and selection can be challenging, because multiple criteria exist for optimal performance. This challenge is particularly great when the sample RNA to be studied comes from fixed, wax-embedded tissue, because such RNA tends to be highly fragmented (K.Specht *et al.*, *Am. J. Pathol* 158: 419-29 [2001]; T.E. Godfrey *et al., J. Mol. Diagnostics* 2:84-91 [2000]).

[0004] It is accepted practice to measure the expression of any given gene by assaying the level of any of its transcribed, spliced, mature mRNA sequences (exon, as opposed to intron, sequence). In theory, an exon is defined as any segment of an interrupted gene that is represented in the mature RNA product, and an intron is defined as a segment of DNA that is transcribed but removed from within the transcript by splicing together the exons on either side of it [B. Lewin. *Genes IV,* Cell Press, Cambridge Mass. 1990]. The rationale for the accepted practice of using exon sequences is theoretically straightforward because the mature RNAs [mRNAs] encode proteins, which define cell phenotypes, whereas intronic RNA is considered to have comparatively

little influence on cell phenotype. Moreover, the prevailing view is that introns are rapidly degraded and therefore more difficult to detect than exon sequences {see introductions of the following articles: Thomas *et al., J. Virol.* 76:532-40 [2002]; Clement *et al., J. Biol. Chem.* 276:16919-30 [2001]; Sharp *et al., Ann. Rev. Biochem.* 55:1119-1150 [1986]}.

[0005] The present invention concerns the use of intronic RNA for measuring gene expression. It will be shown that intronic RNA sequences tend to be readily detected by RT-PCR, even using extensively degraded RNA from fixed tissues. Furthermore, they tend to correlate in their expression with their respective exons. The latter point is particularly unexpected because little or no evidence exists that the ratio of the overall rate constants for synthesis and turnover of transcribed intron and exon sequences are similar. In fact, the scientific literature provides evidence for the complexity of pre-mRNA and spliced intron turnover. For example, pre-mRNA can exist in multiple kinetic pools (Elliott and Rosbash, *Exp. Cell Res.* 229:181-8 [1996]), with subpopulations containing intron RNAs that are not efficiently spliced out and are transported to the cytoplasm as "immature" mRNA species, where they can decay at rates different than nuclear intron RNA sequences (Wang *et al., Proc. Natl. Acad. Sci. USA* 94:4360-5 [1997]). Furthermore, certain spliced intron RNAs seem to enter the cytoplasm in lariat structure (Clement *et al., RNA* 5:206-20 [1999]).

Summary of the Invention

[0006] The present invention is based on experimental evidence demonstrating that transcribed intron sequences, which by definition are present in heterogeneous nuclear RNA but typically are not incorporated into mRNA, have diagnostic and prognostic utility. This is a significant discovery for several reasons. Typically, intron sequences are longer than exon sequences, by twenty fold or more. Thus, introns, given their much greater average length, provide proportionally increased opportunity for optimal gene expression probe design, for example, in the case of RT-PCR, creation of probe/primer sets that possess better technical performance. Independently, because intron sequences evolve more rapidly than exon sequences, intronic RNAs are well-suited to monitor the expression of different closely related members of a gene family.

[0007] In one aspect, the invention concerns a single-stranded oligonucleotide molecule comprising or complementary to a target sequence within a transcribed intronic

RNA sequence of a target gene, wherein the expression of the intronic RNA sequence has been determined to roccelate with the expression of an exonic mRNA sequence within the target gene.

[0008] The single-stranded oligonucleotide molecule can, for example, be a PCR primer or probe. The target sequence typically, but not necessarily, is at least about 55 nucleotide bases long, or at least about 60 nucleotide bases long.

[0009] In an embodiment, the single-straded oligonucleotide molecule is a PCR primer, which is about 17- to 30 nucleotide bases in length.

[0010] In another embodiment, the PCR primer contains about 20% to about 80% G+C bases.

[0011] In yet another embodiment, the PCR primer has a melting temperature (Tm) of between about 50 °C to about 70 °C.

[0012] In a further embodiment, the single-stranded oligonucleotide molecule is a PCR probe, which may be detectably labeled, for example with a reporter fluorescent dye and a quencher fluorescent dye.

[0013] In a further specific embodiment, the target gene is CEGP1, FGXM1, PRAME, or STK15.

[0014] In another specific embodiment, the target gene is selected from the genes listed in Figure 6.

[0015] In a still further specific embodiment, the target gene is selected from the group consisting of B-actin, BAG1, bcl-2, CCNB1, CD68. CEGP1, CTSL2, EstR1, GAPDH, GUS, GRB7, HER2, Ki-67, MYBL2, PR, RPLPO, STK15, STMY3, SURVIVIN, and TFRC.

[0016] In another aspect, the invention concerns a method for monitoring gene expression in a biological sample, comprising:

(a) providing a polynucleotide complementary to an intronic RNA sequence within a target gene, wherein the expression of such intronic RNA sequence correlates with the expression of an exonic mRNA sequence within the target gene;

(b) hybridizing the polynucleotide to the intronic RNA sequence to form a polynucleotide-intronic RNA complex; and

(c) detecting the polynucleotide-intronic RNA complex.

[0017] In a particular aspect, expression of the target gene is measured by RT-PCR, in which case an intron-based primer/probe set can be used in the above process.

[0018] In another aspect, the invention concerns methods of using intron-based sequences to design and create primer-probe sets for RT-PCR. Such primers and probes are particularly suitable to detect and quantify levels of intron RNA in fixed, paraffin-embedded tissue (FPET) specimens, for high sensitivity gene expression analysis. Accordingly, in a further aspect, the invention concerns using intron-based primer-probe sets in gene expression profiling assays, such as gene expression analysis of FPET samples to diagnose and/or predict the prognosis of various pathologic conditions.

[0019] In particular, the invention concerns a method of preparing a single-stranded oligonucleotide molecule for amplification of a target gene, and measuring the level of an intronic RNA species comprising:

(a) identifying at least one intron sequence within the target gene, wherein the expression of the intron sequence correlates with the expression of an exon sequence within the target gene;

(b) preparing a single-stranded oligonucleotide molecule that corresponds to at least a portion of the transcribed intron sequence; and

(c) using the oligonucleotide molecule to measure gene expression.

[0020] Just as before, gene expression can be measured, for example, by RT-PCR, in which case an intron-based primer/probe set (consisting of two primers and a probe) is used to measure gene expression.

[0021] If the oligonucleotide is a forward primer, it is typically designed to comprise 5'-sequences of a target sequence within the transcribed intron sequence. If the oligonucleotide is a reverse primer, it is typically designed to complement 5'-sequences of a target sequence downstream of the forward primer within the transcribed intron sequence. It is important to identify and use a sufficiently long target sequence for PCR amplification. The target sequence generally should be at least about 50 nucleotide bases long, in particular at least 55 nucleotide bases long, in some embodiments at least about 60 nucleotide bases long. The PCR primers and probes are designed following well known principles. Thus, the PCR primer is typically 17-30 nucleotide bases in length, and usually contains about 20% to 80% G+C bases. It is desirable to design PCR primers with a melting temperature (Tm) between about 50°C and about 70°C.

[0022] When the single-stranded oligonucleotide molecule is a PCR probe, it is usually designed to comprise or complement an internal portion of a target sequence within the transcribed intron sequence. For TaqMan® amplification, the PCR probe is labeled with a reporter fluorescent dye and a quencher moiety.

[0023] In another aspect, the invention concerns a method for measuring the expression of a gene by amplifying a target gene by polymerase chain reaction (PCR) comprising:

(a) identifying at least one target intron sequence within the target gene, wherein the expression of the intron sequence correlates with the expression of a corresponding exon sequence within the target gene; and

(b) amplifying the transcribed target intron sequence using an intron-specific PCR primer/probe set.

[0024] The target intron sequence is typically at least about 50 bases long, and the PCR primer and probe set is designed to correspond to unique sequences within the transcribed target intron sequence.

[0025] In yet another aspect, the invention concerns a method for amplifying RNA fragments in a sample representing at least one gene of interest, comprising the steps of:

(a) contacting the sample with at least one set of PCR primers and probe; and

(b) performing PCR amplification,

wherein the PCR primers and probe are designed based upon an intron sequence identified within the gene of interest, and wherein the expression of the intron sequence correlates with the expression of an exon sequence within the gene of interest.

[0026] In particular embodiment, the PCR primers and probe are typically designed based upon a unique sequence within the intron identified. In another embodiment, the sample comprises fragmented RNA representing multiple genes of interest, and is contacted with a pool of PCR primers and probes designed based upon unique sequences within introns present in the genes of interest.

[0027] In a preferred embodiment, the amplification is performed on a fixed, paraffin-embedded tissue (FPET) sample, which can, for example, originate from a tumor biopsy obtained from a human patient. The tumor can be any kind of solid tumor, such as, for example, breast cancer, lung cancer, or colorectal cancer. The tumor tissue can be harvested by a variety of methods, including fine needle biopsy, core biopsy or resection.

[0028] In a particular embodiment, the invention concerns methods using intron-based PCR primer-probe sets in gene expression analysis to predict the likelihood of recurrent disease for patients with early breast cancer.

[0029] In a further aspect, the invention concerns an array comprising a plurality of polynucleotides hybridizing to target genes of interest, wherein preferably at least 70% of the polynucleotides comprises intron sequences.

[0030] In yet another aspect, the invention concerns intron-based amplicon sequences, and their use in gene expression analysis.

[0031] In a particular embodiment the invention concerns gene expression analysis of a biological sample representative of invasive breast cancer based on determining the expression levels of the RNA transcripts or expression products of a gene or gene set selected from the group consisting of:

(a) Bcl2, cyclinG1, NFKBp65, NME1, EPHX1, TOP2B, DR5, TERC, Src, DIABLO;

(b) Ki67, XIAP, hENT1, TS, CD9, p27, cyclinG1, pS2, NFKBp65, CYP3A4;

(c) GSTM1, XIAP, Ki67, TS, cyclinG1, p27, CYP3A4, pS2, NFKBp65, ErbB3;

(d) PR, NME1, XIAP, upa, cyclinG1, Contig51037, TERC, EPHX1, ALDH1A3, CTSL;

(e) CA9, NME1, TERC, cyclinG1, EPHX1, DPYD, Src, TOP2B, NFKBp65, VEGFC;

(f) TFRC, XIAP, Ki67, TS, cyclinG1, p27, CYP3A4, pS2, ErbB3, NFKBp65;

(g) Bcl2, PRAME, cyclinG1, FOXM1, NFKBp65, TS, XIAP, Ki67, CYP3A4, p27;

(h) FOXM1, cyclinG1, XIAP, Contig51037, PRAME, TS, Ki67, PDGFRa, p27, NFKBp65;

(i) PRAME, FOXM1, cyclinG1, XIAP, Contig51037, TS, Ki6, PDGFRa, p27, NFKBp65;

(j) Ki67, XIAP, PRAME, hENT1, contig51037, TS, CD9, p27, ErbB3, cyclinG1;

(k) STK15, XIAP, PRAME, PLAUR, p27, CTSL, CD18, PREP, p53, RPS6KB1;

(l) GSTM1, XIAP, PRAME, p27, Contig51037, ErbB3, GSTp, EREG, ID1, PLAUR;

(m) PR, PRAME, NME1, XIAP, PLAUR, cyclinG1, Contig51037, TERC, EPHX1, DR5;

(n) CA9, FOXM1, cyclinG1, XIAP, TS, Ki67, NFKBp65, CYP3A4, GSTM3, p27;

(o) TFRC, XIAP, PRAME, p27, Contig51037, ErbB3, DPYD, TERC, NME1, VEGFC; and

(p) CEGP1, PRAME, hENT1, XIAP, Contig51037, ErbB3, DPYD, NFKBp65, ID1, TS, including the use of intron-based sequences.

[0032] In another embodiment, the invention concerns gene expression analysis of a biological sample representative of ER-positive breast cancer based on determining the expression levels of the RNA transcripts or expression products of a gene or gene set selected from the group consisting of:

(a) PRAME, p27, IGFBP2, HIF1A, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(b) Contig51037, EPHX1, Ki67, TIMP2, cyclinG1, DPYD, CYP3A4, TP, AIB1, CYP2C8;

(c) Bcl2, hENT1, FOXM1, Contig51037, cyclinG1, Contig46653, PTEN, CYP3A4, TIMP2, AREG;

(d) HIF1A, PRAME, p27, IGFBP2, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(e) IGF1R, PRAME, EPHX1, Contig51037, cyclinG1, Bcl2, NME1, PTEN, TBP, TIMP2;

(f) FOXM1, Contig51037, VEGFC, TBP, HIF1A, DPYD, RAD51C, DCR3, cyclinG1, BAG1;

(g) EPHX1, Contig51037, Ki67, TIMP2, cyclinG1, DPYD, CYP3A4, TP, AIB1, CYP2C8;

(h) Ki67, VEGFC, VDR, GSTM3, p27, upa, ITGA7, rhoC, TERC, Pin1;

(i) CDC25B, Contig51037, hENT1, Bcl2, HLAG, TERC, NME1, upa, ID1, CYP;

(j) VEGFC, Ki67, VDR, GSTM3, p27, upa, ITGA7, rhoC, TERC, Pin1;

(k) CTSB, PRAME, p27, IGFBP2, EPHX1, CTSL, BAD, DR5, DCR3, XIAP;

(l) DIABLO, Ki67, hENT1, TIMP2, ID1, p27, KRT19, IGFBP2, TS, PDGFB;

(m) p27, PRAME, IGFBP2, HIF1A, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;

(n) CDH1; PRAME, VEGFC; HIF1A; DPYD, TIMP2, CYP3A4, EstR1, RBP4, p27;

(o) IGFBP3, PRAME, p27, Bcl2, XIAP, EstR1, Ki67, TS, Src, VEGF;

(p) GSTM3, PRAME, p27, IGFBP3, XIAP, FGF2, hENT1, PTEN, EstR1, APC;

(q) hENT1, Bcl2, FOXM1, Contig51037, CyclinG1, Contig46653, PTEN, CYP3A4, TIMP2, AREG;

(r) STK15, VEGFC, PRAME, p27, GCLC, hENT1, ID1, TIMP2, EstR1, MCP1;

(s) NME1, PRAM, p27, IGFBP3, XIAP, PTEN, hENT1, Bcl2, CYP3A4, HLAG;

(t) VDR, Bcl2, p27, hENT1, p53, PI3KC2A, EIF4E, TFRC, MCM3, ID1;

(u) EIF4E, Contig51037, EPHX1, cyclinG1, Bcl2, DR5, TBP, PTEN, NME1, HER2;

(v) CCNB1, PRAME, VEGFC, HIF1A, hENT1, GCLC, TIMP2, ID1, p27, upa;

(w) ID1, PRAME, DIABLO, hENT1, p27, PDGFRa, NME1, BIN1, BRCA1, TP;

(x). FBXO5, PRAME, IGFBP3, p27, GSTM3, hENT1, XIAP, FGF2, TS, PTEN;

(y) GUS, HIA1A, VEGFC, GSTM3, DPYD, hENT1, EBXO5, CA9, CYP, KRT18; and

(z) Bclx, Bcl2, hENT1, Contig51037, HLAG, CD9, ID1, BRCA1, BIN1, HBEGF, including the use of intron-based sequences.

[0033] In a further embodiment, the cancer is breast cancer, and the gene(s) analyzed is/are selected from the group consisting of: FOXM1; PRAME; SKT15, Ki-67; CA9; NME1; SURV; TFRC; YB-1; RPS6KB1; Src; Chk1; CCNB1; Chk2; CDC25B; CYP3A4; EpCAM; VEGFC; hENT1; BRCA2; EGFR; TK1; VDR; Blc12; CEGP1; GSTM1; PR; BBC3; GATA3; DPYD; GSTM3; ID1; EstR1; p27; XIAP; IGF1R; AK055699; P13KC2A; TGFB3; BAGI1; pS2; WISP1; HNF3A; and NFKBp65.

[0034] In a still further embodiment, invention concerns gene expression analysis of a biological sample representative of invasive breast cancer, based on determining the expression levels of the RNA transcripts or expression products of a gene or gene set selected from the group consisting of:

(a) p53BP2, Bcl2, BAD, EPHX1, PDGFRβ, DIABLO, XIAP, YB1, CA9, and KRT8;

(b) GRB7, CD68, TOP2A, Bcl2, DIABLO, CD3, ID1, PPM1D, MCM6, and WISP1;

(c) PR, p53BP2, PRAME, DIABLO, CTSL, IGFBP2, TIMP1, CA9, MMP9, and COX2;

(d) CD68, GRB7, TOP2A, Bcl2, DIABLO, CD3, ID1, PPM1D, MCM6, and WISP1;

(e) Bcl2, p53BP2, BAD, EPHX1, PDGFRβ, DIABLO, XIAP, YB1, CA9, and KRT8;

(f) KRT14, KRT5, PRAME, p53BP2, GUS1, AIB1, MCM3, CCNE1, MCM6, and ID1;

(g) PRAME, p53BP2, EstR1, DIABLO, CTSL, PPM1D, GRB7, DAPK1, BBC3, and VEGFB;

(h) CTSL2, GRB7, TOP2A, CCNB1, Bcl2, DIABLO, PRAME, EMS1, CA9, and EpCAM;

(i) EstR1, p53BP2, PRAME, DIABLO, CTSL, PPM1D, GRB7, DAPK1, BBC3, and VEGFB;

(j) Chk1, PRAME, p53BP2, GRB7, CA9, CTSL, CCNB1, TOP2A, tumor size, and IGFBP2;

(k) IGFBP2, GRB7, PRAME, DIABLO, CTSL, β-Catenin, PPM1D, Chk1, WISP1, and LOT1;

(l) HER2, p53BP2, Bcl2, DIABLO, TIMP1, EPHX1, TOP2A, TRAIL, CA9, and AREG;

(m) BAG1, p53BP2, PRAME, IL6, CCNB1, PAI1, AREG, tumor size, CA9, and Ki67;

(n) CEGP1, p53BP2, PRAME, DIABLO, Bcl2, COX2, CCNE1, STK15, and AKT2, and FGF18;

(o) STK15, p53BP2, PRAME, IL6, CCNE1, AKT2, DIABLO, cMet, CCNE2, and COX2;

(p) KLK10, EstR1, p53BP2, PRAME, DIABLO, CTSL, PPM1D, GRB7, DAPK1, and BBC3;

(q) AIB1, p53BP2, Bcl2, DIABLO, TIMP1, CD3, p53, CA9, GRB7, and EPHX1

(r) BBC3, GRB7, CD68, PRAME, TOP2A, CCNB1, EPHX1, CTSL GSTM1, and APC;

(s) CD9, GRB7, CD68, TOP2A, Bcl2, CCNB1, CD3, DIABLO, ID1, and PPM1D;

(t) EGFR, KRT14, GRB7, TOP2A, CCNB1, CTSL, Bcl2, TP, KLK10, and CA9;

(u) HIF1α, PR, DIABLO, PRAME, Chk1, AKT2, GRB7, CCNE1, TOP2A, and CCNB1;

(v) MDM2, p53BP2, DIABLO, Bcl2, AIB1, TIMP1, CD3, p53, CA9, and HER2;

(w) MYBL2, p53BP2, PRAME, IL6, Bcl2, DIABLO, CCNE1, EPHX1, TIMP1, and CA9;

(x) p27, p53BP2, PRAME, DIABLO, Bcl2, COX2, CCNE1, STK15, AKT2, and ID1;

(y) RAD51, GRB7, CD68, TOP2A, CIAP2, CCNB1, BAG1, IL6, FGFR1, and p53BP2;

(z) SURV, GRB7, TOP2A, PRAME, CTSL, GSTM1, CCNB1, VDR, CA9; and CCNE2;

(aa) TOP2B, p53BP2, DIABLO, Bcl2, TIMP1, AIB1, CA9, p53, KRT8, and BAD;

(ab) ZNF217, GRB7, p53BP2, PRAME, DIABLO, Bcl2, COX2, CCNE1, APC4, and β-Catenin.

[0035] In a different embodiment, the invention concerns gene expression analysis of a biological sample, using intron-based polynucleotide sequences hybridizing to at least one genes selected from the group consisting of: CD68; CTSL; FBXO5; SURV; CCNB1; MCM2; Chk1; MYBL2; HIF1A; cMET; EGFR; TS; STK15, IGFR1; BCl2; HNF3A; TP53BP2; GATA3; BBC3; RAD51C; BAG1; IGFBP2; PR; CD9; RB1; EPHX1; CEGP1; TRAIL; DR5; p27; p53; MTA; RIZ1; ErbB3; TOP2B; EIF4E, CD68; CTSL; FBXO5; SURV; CCNB1; MCM2; Chk1; MYBL2; HIF1A; cMET; EGFR; TS; and STK15.

[0036] Gene expression analysis may be performed in an array format, and the array preferably is a high-density array, comprising at least 100, more preferably at least 150, even more preferably, 200 sequences in a 5-10 μ section.

Brief Description of the Drawings

[0037] Figures 1A-M show masked intron sequences for the CEGP1, FOXM1, PRAME, and STK15 genes. Amplicons used for RT-PCR are shown in italics.

[0038] Figure 2 shows primer/probe sets for CEGP1, FOXM1, PRAME, and STK15. Sequences of forward and reverse primers are indicated by "F" and "R," respectively. Sequences of primers are designated with "P."

[0039] Figure 3 shows correlation coefficients [R] for co-expression of CEGP1 exon RNA with 47 other RNA sequences. Symbols: diamond = CEGP1 exon self vs. self (=1.0 by definition); squares=CEGP1 introns; triangles = sequences of other genes.

[0040] Figure 4 shows correlation coefficients [R] for co-expression of PRAME exon RNA with 47 other RNA sequences. Symbols: diamond – PRAME exon self vs. self (=1.0 by definition); squares=PRAME introns; triangles = sequences of other genes.

[0041] Figure 5 shows correlation coefficients [R] for co-expression of STK15 exon RNA with 47 other RNA sequences. Symbols: diamond – STK15 exon self vs. self (=1.0 by definition); squares=STK15 introns; triangles = sequences of other genes.

[0042] Figure 6 shows an exemplary set of genes, the expression of which can be analyzed by the methods of the present invention.

Detailed Description of the Preferred Embodiment

A. Definitions

[0043] Unless defined otherwise, technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Singleton *et al.*, Dictionary of Microbiology and Molecular Biology 2nd ed., J. Wiley & Sons (New York, NY 1994), provide one skilled in the art with a general guide to many of the terms used in the present application.

[0044] One skilled in the art will recognize many methods and materials similar or equivalent to those described herein, which could be used in the practice of the present invention. Indeed, the present invention is in no way limited to the methods and materials described. For purposes of the present invention, the following terms are defined below.

[0045] The terms "splicing" and "RNA splicing" are used interchangeably and refer to RNA processing that removes introns and joins exons to produce mature mRNA with continuous coding sequence that moves into the cytoplasm of an eukaryotic cell.

[0046] In theory, the term "exon" refers to any segment of an interrupted gene that is represented in the mature RNA product (B. Lewin. *Genes IV* Cell Press, Cambridge Mass. 1990). In theory the term "intron" refers to any segment of DNA that is transcribed but removed from within the transcript by splicing together the exons on either side of it. Operationally, exon sequences occur in the mRNA sequence of a gene as defined by Ref. Seq ID numbers. Operationally, intron sequences are the intervening sequences within the genomic DNA of a gene, bracketed by exon sequences and having GT and AG splice consensus sequences at their 5' and 3' boundaries.

[0047] The term "microarray" refers to an ordered arrangement of hybridizable array elements, preferably polynucleotide probes, on a substrate.

[0048] The term "polynucleotide," when used in singular or plural, generally refers to any polyribonucleotide or polydeoxribonucleotide, which may be unmodified RNA or DNA or modified RNA or DNA. Thus, for instance, polynucleotides as defined herein include, without limitation, single- and double-stranded DNA, DNA including single- and double-stranded regions, single- and double-stranded RNA, and RNA including single- and double-stranded regions, hybrid molecules comprising DNA and RNA that may be single-stranded or, more typically, double-stranded or include single- and double-stranded regions. In addition, the term "polynucleotide" as used herein refers to triple-stranded regions comprising RNA or DNA or both RNA and DNA. The strands in such regions may be from the same molecule or from different molecules. The regions may include all of one or more of the molecules, but more typically involve only a region of some of the molecules. One of the molecules of a triple-helical region often is an oligonucleotide. The term "polynucleotide" specifically includes cDNAs. The term includes DNAs (including cDNAs) and RNAs that contain one or more modified bases. Thus, DNAs or RNAs with backbones modified for stability or for other reasons are "polynucleotides" as that term is intended herein. Moreover, DNAs or RNAs comprising unusual bases, such as inosine, or modified bases, such as tritiated bases, are included within the term "polynucleotides" as defined herein. In general, the term "polynucleotide" embraces all chemically, enzymatically and/or metabolically modified

forms of unmodified polynucleotides, as well as the chemical forms of DNA and RNA characteristic of viruses and cells, including simple and complex cells.

[0049] The term "oligonucleotide" refers to a relatively short polynucleotide, including, without limitation, single-stranded deoxyribonucleotides, single- or double-stranded ribonucleotides, RNA:DNA hybrids and double-stranded DNAs. Oligonucleotides, such as single-stranded DNA probe oligonucleotides, are often synthesized by chemical methods, for example using automated oligonucleotide synthesizers that are commercially available. However, oligonucleotides can be made by a variety of other methods, including *in vitro* recombinant DNA-mediated techniques and by expression of DNAs in cells and organisms.

[0050] The terms "differentially expressed gene," "differential gene expression" and their synonyms, which are used interchangeably, refer to a gene whose expression is at a higher or lower level in one patient or test subject relative to another, for example, in a subject suffering from a disease, specifically cancer, such as breast cancer, relative to its expression in a normal or control subject. The terms also include genes whose expression is activated to a higher or lower level at different stages of the same disease. It is also understood that a differentially expressed gene may be either activated or inhibited at the nucleic acid level or protein level, or may be subject to alternative splicing to result in a different polypeptide product. Such differences may be evidenced by a change in mRNA levels, surface expression, secretion or other partitioning of a polypeptide, for example. Differential gene expression may include a comparison of expression between two or more genes or their gene products, or a comparison of the ratios of the expression between two or more genes or their gene products, or even a comparison of two differently processed products of the same gene, which differ between normal subjects and subjects suffering from a disease, specifically cancer, or between various stages of the same disease. Differential expression includes both quantitative, as well as qualitative, differences in the temporal or cellular expression pattern in a gene or its expression products among, for example, normal and diseased cells, or among cells which have undergone different disease events or disease stages. For the purpose of this invention, "differential gene expression" is considered to be present when there is at least an about two-fold, preferably at least about four-fold, more preferably at least about six-fold, most preferably at least about ten-fold difference between the expression of a given gene in normal and diseased subjects, or in various stages of disease development in a diseased subject.

[0051] The term "normalized" with regard to a gene transcript or a gene expression product refers to the level of the transcript or gene expression product relative to the mean levels of transcripts/products of a set of reference genes, wherein the reference genes are either selected based on their minimal variation across, patients, tissues or treatments ("housekeeping genes"), or the reference genes are the totality of tested genes. In the latter case, which is commonly referred to as "global normalization", it is important that the total number of tested genes be relatively large, preferably greater than 50. Specifically, the term 'normalized' with respect to an RNA transcript refers to the transcript level relative to the mean of transcript levels of a set of reference genes. More specifically, the mean level of an RNA transcript as measured by TaqMan® RT-PCR refers to the Ct value minus the mean Ct values of a set of reference gene transcripts.

[0052] The term "over-expression" with regard to an RNA transcript is used to refer to the level of the transcript determined by normalization to the level of reference mRNAs, which might be all measured transcripts in the specimen or a particular reference set of mRNAs.

[0053] The terms "expression threshold," and "defined expression threshold" are used interchangeably and refer to the level of a gene or gene product in question above which the gene or gene product serves as a predictive marker for patient response or resistance to a drug,. The threshold typically is defined experimentally from clinical studies. The expression threshold can be selected either for maximum sensitivity (for example, to detect all responders to a drug), or for maximum selectivity (for example to detect only responders to a drug), or for minimum error.

[0054] The phrase "gene amplification" refers to a process by which multiple copies of a gene or gene fragment are formed in a particular cell or cell line. The duplicated region (a stretch of amplified DNA) is often referred to as "amplicon." Often, the amount of the messenger RNA (mRNA) produced, *i.e.,* the level of gene expression, also increases in the proportion of the number of copies made of the particular gene expressed.

[0055] The term "prognosis" is used herein to refer to the prediction of the likelihood of cancer-attributable death or progression, including recurrence, metastatic spread, and drug resistance, of a neoplastic disease, such as breast cancer. The term "prediction" is used herein to refer to the likelihood that a patient will respond either favorably or unfavorably to a drug or set of drugs, and also the extent of those responses, or that a patient will survive, following surgical removal or the primary tumor and/or

chemotherapy for a certain period of time without cancer recurrence. The predictive methods of the present invention can be used clinically to make treatment decisions by choosing the most appropriate treatment modalities for any particular patient. The predictive methods of the present invention are valuable tools in predicting if a patient is likely to respond favorably to a treatment regimen, such as surgical intervention, chemotherapy with a given drug or drug combination, and/or radiation therapy, or whether long-term survival of the patient, following sugery and/or termination of chemotherapy or other treatment modalities is likely.

[0056] The term "long-term" survival is used herein to refer to survival for at least 3 years, more preferably for at least 5 years, most preferably for at least 10 years following surgery or other treatment.

[0057] The term "increased resistance" to a particular drug or treatment option, when used in accordance with the present invention, means decreased response to a standard dose of the drug or to a standard treatment protocol.

[0058] The term "decreased sensitivity" to a particular drug or treatment option, when used in accordance with the present invention, means decreased response to a standard dose of the drug or to a standard treatment protocol, where decreased response can be compensated for (at least partially) by increasing the dose of drug, or the intensity of treatment.

[0059] "Patient response" can be assessed using any endpoint indicating a benefit to the patient, including, without limitation, (1) inhibition, to some extent, of tumor growth, including slowing down and complete growth arrest; (2) reduction in the number of tumor cells; (3) reduction in tumor size; (4) inhibition (i.e., reduction, slowing down or complete stopping) of tumor cell infiltration into adjacent peripheral organs and/or tissues; (5) inhibition (i.e. reduction, slowing down or complete stopping) of metastasis; (6) enhancement of anti-tumor immune response, which may, but does not have to, result in the regression or rejection of the tumor; (7) relief, to some extent, of one or more symptoms associated with the tumor; (8) increase in the length of survival following treatment; and/or (9) decreased mortality at a given point of time following treatment.

[0060] The term "treatment" refers to both therapeutic treatment and prophylactic or preventative measures, wherein the object is to prevent or slow down (lessen) the targeted pathologic condition or disorder. Those in need of treatment include those already with the disorder as well as those prone to have the disorder or those in

whom the disorder is to be prevented. In tumor (*e.g.*, cancer) treatment, a therapeutic agent may directly decrease the pathology of tumor cells, or render the tumor cells more susceptible to treatment by other therapeutic agents, *e.g.*, radiation and/or chemotherapy.

[0061] The term "tumor," as used herein, refers to all neoplastic cell growth and proliferation, whether malignant or benign, and all pre-cancerous and cancerous cells and tissues.

[0062] The terms "cancer" and "cancerous" refer to or describe the physiological condition in mammals that is typically characterized by unregulated cell growth. Examples of cancer include but are not limited to, breast cancer, colon cancer, lung cancer, prostate cancer, hepatocellular cancer, gastric cancer, pancreatic cancer, cervical cancer, ovarian cancer, liver cancer, bladder cancer, cancer of the urinary tract, thyroid cancer, renal cancer, carcinoma, melanoma, and brain cancer.

[0063] The "pathology" of cancer includes all phenomena that compromise the well-being of the patient. This includes, without limitation, abnormal or uncontrollable cell growth, metastasis, interference with the normal functioning of neighboring cells, release of cytokines or other secretory products at abnormal levels, suppression or aggravation of inflammatory or immunological response, neoplasia, premalignancy, malignancy, invasion of surrounding or distant tissues or organs, such as lymph nodes, etc.

[0064] "Stringency" of hybridization reactions is readily determinable by one of ordinary skill in the art, and generally is an empirical calculation, dependent upon probe length, washing temperature, and salt concentration. In general, longer probes require higher temperatures for proper annealing, while shorter probes need lower temperatures. Hybridization generally depends on the ability of denatured DNA to reanneal when complementary strands are present in an environment below their melting temperature. The higher the degree of desired homology between the probe and hybridizable sequence, the higher the relative temperature which can be used. As a result, it follows that higher relative temperatures would tend to make the reaction conditions more stringent, while lower temperatures less so. For additional details and explanation of stringency of hybridization reactions, see Ausubel et al., Current Protocols in Molecular Biology, Wiley Interscience Publishers, (1995).

[0065] "Stringent conditions" or "high stringency conditions", as defined herein, typically: (1) employ low ionic strength and high temperature for washing, for example 0.015 M sodium chloride/0.0015 M sodium citrate/0.1% sodium dodecyl sulfate

at 50°C; (2) employ during hybridization a denaturing agent, such as formamide, for example, 50% (v/v) formamide with 0.1% bovine serum albumin/0.1% Ficoll/0.1% polyvinylpyrrolidone/50mM sodium phosphate buffer at pH 6.5 with 750 mM sodium chloride, 75 mM sodium citrate at 42°C; or (3) employ 50% formamide, 5 x SSC (0.75 M NaCl, 0.075 M sodium citrate), 50 mM sodium phosphate (pH 6.8), 0.1% sodium pyrophosphate, 5 x Denhardt's solution, sonicated salmon sperm DNA (50 μg/ml), 0.1% SDS, and 10% dextran sulfate at 42°C, with washes at 42°C in 0.2 x SSC (sodium chloride/sodium citrate) and 50% formamide at 55°C, followed by a high-stringency wash consisting of 0.1 x SSC containing EDTA at 55°C.

[0066] "Moderately stringent conditions" may be identified as described by Sambrook et al., Molecular Cloning: A Laboratory Manual, New York: Cold Spring Harbor Press, 1989, and include the use of washing solution and hybridization conditions (e.g., temperature, ionic strength and %SDS) less stringent that those described above. An example of moderately stringent conditions is overnight incubation at 37°C in a solution comprising: 20% formamide, 5 x SSC (150 mM NaCl, 15 mM trisodium citrate), 50 mM sodium phosphate (pH 7.6), 5 x Denhardt's solution, 10% dextran sulfate, and 20 mg/ml denatured sheared salmon sperm DNA, followed by washing the filters in 1 x SSC at about 37-50°C. The skilled artisan will recognize how to adjust the temperature, ionic strength, etc. as necessary to accommodate factors such as probe length and the like.

[0067] In the context of the present invention, reference to "at least one," "at least two," "at least five," etc. of the genes listed in any particular gene set means any one or any and all combinations of the genes listed.

[0068] The term "housekeeping gene" refers to a group of genes that codes for proteins whose activities are essential for the maintenance of cell function. These genes are typically similarly expressed in all cell types. Housekeeping genes include, without limitation, glyceraldehyde-3-phosphate dehydrogenase (GAPDH), Cyp1, albumin, actins, e.g. β-actin, tubulins, cyclophilin, hypoxantine phsophoribosyltransferase (HRPT), L32. 28S, and 18S.

[0069] According to the present invention, a polynucleotide or oligonucleotide molecule "corresponds to" a target sequence, such as an intron sequence or transcribed intronic RNA, if it incorporates or is complementary to such sequence.

B. Detailed Description

[0070] The practice of the present invention will employ, unless otherwise indicated, conventional techniques of molecular biology (including recombinant techniques), microbiology, cell biology, and biochemistry, which are within the skill of the art. Such techniques are explained fully in the literature, such as, "Molecular Cloning: A Laboratory Manual", 2nd edition (Sambrook et al., 1989); "Oligonucleotide Synthesis" (M.J. Gait, ed., 1984); "Animal Cell Culture" (R.I. Freshney, ed., 1987); "Methods in Enzymology" (Academic Press, Inc.); "Handbook of Experimental Immunology", 4th edition (D.M. Weir & C.C. Blackwell, eds., Blackwell Science Inc., 1987); "Gene Transfer Vectors for Mammalian Cells" (J.M. Miller & M.P. Calos, eds., 1987); "Current Protocols in Molecular Biology" (F.M. Ausubel et al., eds., 1987); and "PCR: The Polymerase Chain Reaction", (Mullis et al., eds., 1994).

1. *Polymerase Chain Reaction (PCR)*

[0071] The purpose of the polymerase chain reaction (PCR) is to make copies of a gene in order to provide larger amounts of nucleic acid for further use. PCR is a process based on a specialized polymerase enzyme (e.g. Taq DNA polymerase), which can synthesize a complementary strand to a given DNA strand in a mixture containing the four dNTP's (sATP, dCTP, dGTP, dTTP) and two oligonucleotide primers flanking the target sequence to be amplified. The two oligonucleotide primers are used to generate an amplicon typical of a PCR reaction. A third oligonucleotide, or probe, is designed to detect the nucleotide sequence located between the two PCR primers. Although the probe design might differ, in the TaqMan® PCR method probe signals are controlled by the proximity of a reporter fluorescent dye and a quencher fluorescent dye. Any laser-induced emission from the reporter dye is quenched by the quenching dye when the two dyes are located close together as they are on the probe. During the amplification reaction, the polymerase enzyme (e.g. Taq DNA polymerase) cleaves the probe in a template-dependent manner. The resultant probe fragments disassociate in solution, and signal from the released reporter dye is free from the quenching effect of the second fluorophore. One molecule of reporter dye is liberated for each new molecule synthesized, and detection of the unquenched reporter dye provides the basis for quantitative interpretation of the data.

[0072] The starting material for PCR can be DNA, cDNA, mRNA or any other polynucleotide that needs to be amplified. Since the PCR requires single-stranded

DNA as template, if the starting material is double-stranded DNA, it needs to be denatured in order to produce single-stranded DNA.

[0073] As RNA cannot serve as a template for PCR, if the starting material is RNA, the first step is the reverse transcription of the RNA template into cDNA, followed by its exponential amplification in a PCR reaction. This version of PCR is generally referred to as reverse transcriptase PCR (RT-PCR). The two most commonly used reverse transcriptases are avian myeloblastosis virus reverse transcriptase (AMV-RT) and Moloney murine leukemia virus reverse transcriptase (MMLV-RT). The reverse transcription step is typically primed using specific primers, random hexamers, or oligo-dT primers, depending on the circumstances. For example, RNA extracted from a tissue sample (e.g. FPET) can be reverse-transcribed using a GeneAmp RNA PCR kit (Perkin Elmer, CA, USA), following the manufacturer's instructions. The derived cDNA can then be used as a template in the subsequent PCR reaction.

[0074] Although the PCR step can use a variety of thermostable DNA-dependent DNA polymerases, it typically employs the Taq DNA polymerase, which has a 5'-3' nuclease activity but lacks a 3'-5' proofreading endonuclease activity. Thus, TaqMan® PCR typically utilizes the 5'-nuclease activity of Taq or Tth polymerase to hydrolyze a hybridization probe bound to its target amplicon, but any enzyme with equivalent 5' nuclease activity can be used. In this case, the probe is designed to be non-extendible by Taq DNA polymerase enzyme. TaqMan® RT-PCR can be performed using commercially available equipment, such as, for example, ABI PRISM 7700™ Sequence Detection System™ (Perkin-Elmer-Applied Biosystems, Foster City, CA, USA), or Lightcycler (Roche Molecular Biochemicals, Mannheim, Germany). In a preferred embodiment, the 5' nuclease procedure is run on a real-time quantitative PCR device such as the ABI PRISM 7700™ Sequence Detection System™. The system consists of a thermocycler, laser, charge-coupled device (CCD), camera and computer. The system amplifies samples in a 96-well format on a thermocycler. During amplification, laser-induced fluorescent signal is collected in real-time through fiber optics cables for all 96 wells, and detected at the CCD. The system includes software for running the instrument and for analyzing the data.

[0075] 5'-Nuclease assay data are initially expressed as Ct, or the threshold cycle. As discussed above, fluorescence values are recorded during every cycle and represent the amount of product amplified to that point in the amplification reaction. The

point when the fluorescence signal is first recorded as statistically significant is the threshold cycle (C_t).

[0076] To minimize errors and the effect of sample-to-sample variation, RT-PCR is usually performed using an internal standard. The ideal internal standard is expressed at a constant level among different tissues, and is unaffected by the experimental treatment. RNAs frequently used to normalize patterns of gene expression are mRNAs for the housekeeping genes glyceraldehyde-3-phosphate-dehydrogenase (GAPDH) and β-actin.

[0077] For further details of real time quantitative PCR see also Held *et al., Genome Research* 6:986-994 (1996). PCR is described in U.S. Patent Nos. 4,683,202, 4,683,195; 4,965,188; and 5,075,216, the entire disclosures of which are hereby expressly incorporated by reference.

2. *Introns and RNA Splicing*

[0078] Most genes in higher eukaryotes contain more than 100,000 nucleotide pairs, some containing more than 2 million nucleotide pairs. This is significantly longer than the nucleotide sequence required to encode an average size protein (300-400 amino acids), which is in the order of about 1000 nucleotides. Most of the extra length consists of noncoding (intron) sequences that interrupt the coding (exon) sequences within the gene sequence. Most of higher eukaryotic genes coding for mRNA, tRNA and some coding for rRNA are interrupted by intron sequences. Genes for mRNA typically have 0 to 60 introns; while genes for tRNA typically include 0 or 1 intron.

[0079] When mRNA is transcribed from DNA, at first both exon and intron sequences are transcribed into the so-called heterogeneous nuclear RNA (hnRNA) or immature RNA or pre-mRNA. However, before the RNA exits the nucleus, intron sequences are often deleted from the transcribed mRNA as a result of a process known as RNA splicing. The process of intron removal involves a precise looping process controlled by a specific nucleotide sequence abutting the exons. Almost all introns can be identified by specific consensus sequences. The first two bases of an intron are always GU, while the last two bases are always AG, but the 5' and 3' splice sites typically have consensus sequences that extend beyond the GU and AG motifs. Splicing of mRNA takes place on a particle called spliceosome, while tRNA and rRNA are spliced by mechanisms that do not involve spliceosomes.

[0080] Introns are typically much longer than exons (sequences that are present in the mRNA). An average eukaryotic exon is about 150 nucleotides long, while a single human intron can be as long as close to 500,000 nucleotides, but typically are about 2000-4000 nucleotides. In general, a eukaryotic gene contains much more intron than exon sequences, as illustrated by the following table (Molecular Biology of the Cell, Bruce Alberts et al., eds., 3rd edition, Garland Publishing Company, New York, N.Y., 1994, p. 340):

Table 1

Gene	Gene Size ($\times 10^3$ nucleotides)	MRNA Size ($\times 10^3$ nucleotides)	Number of Introns
β-globin	1.5	0.6	2
Insulin	1.7	0.4	2
Proteinase C	11	1.4	7
Albumin	25	2.1	14
Catalase	34	1.6	12
LDL receptor	45	5.5	17
Factor VIII	186	9	25
Thyroglobulin	300	8.7	36

[0081] In a particular embodiment of the present invention, intron sequences within a gene of interest are subjected to a selection process to identify intronic RNA sequence or sequences that co-express with exon RNA (i.e., mRNA) sequences of the same gene. Such selected intron sequences, the expression of which correlates with the expression of exon sequences, have especially desirable properties as potential diagnostic markers: (1) because of their favorable technical performance (specifically, optimizing assay specificity and sensitivity); and, (2) whatever biomedical importance attaches to the mRNA level of the gene is also attached to the cellular levels of intronic sequences. For example, high levels of an mRNA species that encodes a potent growth factor are likely to correlate with high rate of growth of a cell. Intronic sequences having cellular levels that correlate with mRNA levels of this same gene have the same likelihood to correlate with high growth rate of a cell. Such selected intronic sequences can then be used to screen valuable tissue specimens to search for clinical correlations and diagnostic, predictive or prognostic significance.

[0082] An exemplary process for selecting intron sequences that co-express with the mRNA of the same gene is as follows. Briefly, for any gene of interest, a set of relevant tissues from a population of patients of interest are assayed to measure the levels of a set of intronic and mRNA sequences. The intronic sequences found to have the highest Pearson correlation coefficient for co-expression with exon RNA (mRNA) sequences are then selected. The number of patients studied in this process is preferably at least above 50 and more preferably at least about 100.

[0083] In a specific example, the biomedical issue of interest regards patients with breast cancer and the gene of interest can be the tumor growth marker Ki-67. In this case, tumors from 50 or more breast cancer patients are used for measurement of Ki-67 mRNA levels and the levels of sequences from multiple Ki-67 introns, and the introns having the highest Pearson correlation coefficient for co-expressing with exon RNA are selected.

[0084] An advantage of this process is that the selection of the preferred intronic sequence can be carried out with tissue specimens that are relatively easily obtained and abundant (for example, specimens that lack valuable attached clinical records). Because such tissue can provide large amounts of RNA to screen, it will be possible to detect gene expression signals from even suboptimal probes. The highly sensitive and specific assays based on the selected intronic sequences then can be used to screen valuable tissue specimens, for example, specimens attached to important clinical information, such as disease recurrence, death, or response to defined therapeutic drugs or treatment regimens.

3. *Gene Expression Profiling Using Intron-based PCR Primer/Probe Sets*

[0085] At present, PCR primers and probes are designed based upon the mRNA or cDNA sequence, without considering the intron sequences. Indeed, introns are usually regarded as "packaging" material that is removed during splicing and. generally rapidly degraded.

[0086] The present invention is based on the unanticipated experimental finding that intron RNAs can be readily detected by RT-PCR, even using highly degraded RNA from fixed, paraffin-embedded tissue specimens. In particular, it has been found that in gene expression profiling for a given gene RT-PCR signals from intron-based probe/primer sets can be as large, or larger, than the signals from exon-based RT-PCR signals. While this finding is supported by a few recent findings with certain mRNA

species, it is not in accord with the prevailing view that introns are very rapidly degraded following splicing (Thomas *et al., J. Virol.* 76:532-40 [2002]; Clement *et al., J. Biol. Chem.* 276:16919-30 [2001]; Sharp *et al., Ann. Rev. Biochem.* 55:1119-1150 [1986])..

[0087] Also unexpectedly, the experimental findings underlying the present invention indicate that intronic RNA can be used for gene expression profiling, because the tissue amounts of expressed intron and exon sequences tend to be correlated. This result is unanticipated because scant or no evidence exists that the ratio of the overall rate constants for synthesis and turnover of transcribed intron and exon sequences are similar. In fact, the scientific literature provides evidence for the complexity of pre-mRNA and spliced intron turnover. For example, pre-mRNA can exist in multiple kinetic pools (Elliott and Rosbash, *Exp. Cell Res.* 229:181-8 [1996]), with subpopulations containing intron RNAs that are not efficiently spliced out and are transported to the cytoplasm in "immature" mRNA species, where they can decay at rates different than nuclear intron RNA sequences (Wang *et al., Proc. Natl. Acad. Sci. USA* 94:4360-5 [1997]). Evidence exists that certain spliced intron RNAs enter the cytoplasm in lariat structure (Clement *et al., RNA* 5:206-20 [1999]).

[0088] Finally, data presented here indicate that intron sequences can serve as diagnostic or prognostic molecular markers. Examining four mRNAs previously demonstrated to be prognostic in cancer, it is shown that their corresponding intron sequences are also prognostic, and in the same directions as the parent transcribed exon sequences (i.e., either positively or negatively prognostic).

[0089] In brief, the approach of the invention has been demonstrated as follows. Co-pending application Serial No. 10/388,360, filed on March 12, 2003 (PCT/US03/07713), the entire disclosure of which is hereby expressly incorporated by reference, describes a set of genes that predict likelihood of breast cancer recurrence. In that study, the levels of transcribed exon sequences in fixed paraffin-embedded breast cancer tissue specimens from 146 patients were measured by RT-PCR using exon-based PCR primer/probe sets. In the study described here, RT-PCR assays were created to measure the levels of transcribed intron sequences within four of the previously identified marker genes, and then used to screen RNA from 60 fixed paraffin-embedded biopsy specimens (representing 60 different patients, a subset of the patients evaluated in the previous study). The data presented in the examples below show that for each gene the introns and exons are co-expressed, and that the introns predict risk of disease recurrence as predicted by the previous exon-based data.

4. *Design of Intron-Based PCR Primers and Probes*

[0090] According to one aspect of the present invention, PCR primers and probes are designed based upon intron sequences present in the gene to be amplified. Accordingly, the first step in the primer/probe design is the delineation of intron sequences within the genes. This can be done by publicly available software, such as the DNA BLAT software developed by Kent, W.J., *Genome Res.* 12(4):656-64 (2002), or by the BLAST software including its variations. Subsequent steps follow well established methods of PCR primer and probe design.

[0091] In order to avoid non-specific signals, it is important to mask repetitive sequences within the introns when designing the primers and probes. This can be easily accomplished by using the Repeat Masker program available on-line through the Baylor College of Medicine, which screens DNA sequences against a library of repetitive elements and returns a query sequence in which the repetitive elements are masked. The masked intron sequences can then be used to design primer and probe sequences using any commercially or otherwise publicly available primer/probe design packages, such as Primer Express (Applied Biosystems); MGB assay-by–design (Applied Biosystems); Primer3 (Steve Rozen and Helen J. Skaletsky (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) *Bioinformatics Methods and Protocols: Methods in Molecular Biology.* Humana Press, Totowa, NJ, pp 365-386).

[0092] The most important factors considered in PCR primer design include primer length, melting temperature (Tm), and G/C content, specificity, complementary primer sequences, and 3'-end sequence. In general, optimal PCR primers are generally 17-30 bases in length, and contain about 20-80%, such as, for example, about 50-60% G+C bases. Tm's between 50 and 80 °C, e.g. about 50 to 70 °C are typically preferred.

[0093] For further guidelines for PCR primer and probe design see, e.g. Dieffenbach, C.W. *et al.,* "General Concepts for PCR Primer Design" in: *PCR Primer, A Laboratory Manual,* Cold Spring Harbor Laboratory Press, New York, 1995, pp. 133-155; Innis and Gelfand, "Optimization of PCRs" in: *PCR Protocols, A Guide to Methods and Applications,* CRC Press, London, 1994, pp. 5-11; and Plasterer, T.N. Primerselect: Primer and probe design. *Methods Mol. Biol.* 70:520-527 (1997), the entire disclosures of which are hereby expressly incorporated by reference.

5. *Applications*

[0094] The methods of the present invention, and specifically, the intron-based PCR primers and probes herein, utility in all fields where amplification of a nucleic acid (including RNA, DNA and, in general, all oligo- and poly nucleotides) representing a gene or a gene fragment is required. Thus the PCR primers and probes designed in accordance with the present invention can be used to amplify individual genes, or multiple genes present in a biological sample for the purpose of gene expression profiling by any methodology including, without limitation, gene expression profiling relying on quantitative PCR (e.g. quantitative RT-PCR), and microarray analysis, and bead-based assays.

[0095] For example, in a specific embodiment of the microarray technique, PCR amplified inserts of cDNA clones are applied to a substrate in a dense array. Preferably at least 10,000 nucleotide sequences are applied to the substrate. The microarrayed genes, immobilized on the microchip at at least 10,000 elements each, are suitable for hybridization under stringent conditions. Fluorescently labeled cDNA probes may be generated through incorporation of fluorescent nucleotides by reverse transcription of RNA extracted from tissues of interest. Labeled cDNA probes applied to the chip hybridize with specificity to each spot of DNA on the array. After stringent washing to remove non-specifically bound probes, the chip is scanned by confocal laser microscopy or by another detection method, such as a CCD camera. Quantitation of hybridization of each arrayed element allows for assessment of corresponding mRNA abundance. With dual color fluorescence, separately labeled cDNA probes generated from two sources of RNA are hybridized pairwise to the array. The relative abundance of the transcripts from the two sources corresponding to each specified gene is thus determined simultaneously. The miniaturized scale of the hybridization affords a convenient and rapid evaluation of the expression pattern for large numbers of genes. Such methods have been shown to have the sensitivity required to detect rare transcripts, which are expressed at a few copies per cell, and to reproducibly detect at least approximately two-fold differences in the expression levels (Schena *et al., Proc. Natl. Acad. Sci. USA* 93(2):106-149 (1996)). Microarray analysis can be performed by commercially available equipment, following manufacturer's protocols, such as by using the Affymetrix GenChip technology, or Agilent's microarray technology.

[0096] An important aspect of the present invention is to use intron-based gene amplification as part of gene expression profiling to match patients to best drugs or

drug combinations, and to provide prognostic information. For example, the measured expression of genes in cancer tissue (e.g. biopsied breast cancer tissue) can be used to predict the likelihood of long-term, disease-free survival of patients following surgery and/or other cancer therapy, or to predict patient response to a particular therapeutic approach. For this purpose it is typically necessary to correct for (normalize away) both differences in the amount of RNA assayed and variability in the quality of the RNA used. Therefore, the assays of the invention usually measure and incorporate the expression of certain normalizing genes, including well known reference genes, such as GAPDH and Cyp1. Alternatively, normalization can be based on the mean or median signal (Ct) of all of the assayed genes or a large subset thereof (global normalization approach). On a gene-by-gene basis, the measured normalized amount of a patient tumor mRNA is compared to the amount found in a cancer, e.g. breast cancer tissue reference set. The number (N) of cancer, e.g. breast cancer, tissues in this reference set should be sufficiently high to ensure that different reference sets (as a whole) behave essentially the same way. If this condition is met, the identity of the individual breast cancer tissues present in a particular set will have no significant impact on the relative amounts of the genes assayed. Usually, the breast cancer tissue reference set consists of at least about 30, preferably at least about 40 different fixed, paraffin-embedded (FPE) breast cancer tissue specimens. Unless noted otherwise, normalized expression levels for each mRNA/tested tumor/patient will be expressed as a percentage of the expression level measured in the reference set. More specifically, the reference set of a sufficiently high number (e.g., 40) tumors yields a distribution of normalized levels of each mRNA species. The level measured in a particular tumor sample to be analyzed falls at some percentile within this range, which can be determined by methods well known in the art.

[0097] In a Phase II study of gene expression in paraffin-embedded, fixed tissue samples of invasive breast carcinoma, the overexpression of any of the following genes in the breast cancer tissue was found to indicate a reduced likelihood of survival without cancer recurrence following surgery: FOXM1; PRAME; SKT15, Ki-67; CA9; NME1; SURV; TFRC; YB-1; RPS6KB1; Src; Chk1; CCNB1; Chk2; CDC25B; CYP3A4; EpCAM; VEGFC; hENT1; BRCA2; EGFR; TK1; VDR.

[0098] In the same study, the overexpression of any of the following genes in breast cancer indicates a better prognosis for survival without cancer recurrence following surgery: Blc12; CEGP1; GSTM1; PR; BBC3; GATA3; DPYD; GSTM3; ID1; EstR1;

p27; XIAP; IGF1R; AK055699; P13KC2A; TGFB3; BAGI1; pS2; WISP1; HNF3A; NFKBp65.

[0099] In this same Phase II study of gene expression in paraffin-embedded, fixed tissue samples of ER-positive breast cancer, overexpression of the following genes was indicative of a reduced likelihood of survival without cancer recurrence following surgery: PRAME; FOXM1; EPHX1; HIF1A; VEGFC; Ki-67; VDR; NME1. Some of these genes (PRAME; FOXM1; VEGFC; Ki-67; VDR; and NME1) were also identified as indicators of poor prognosis in the previous analysis, not limited to ER-positive breast cancer. The overexpression of the remaining genes (EPHX1 and HIF1A) was found to be negative indicator of disease free survival in ER-positive breast cancer only. Overexpression of the following genes in ER-positive cancer was found to be indicative of a better prognosis for survival without cancer recurrence following surgery: Bcl-2; DIABLO; IGF1R; GSTM3. Of the latter genes, Bcl-2; IGFR1; and GSTM3 have also been identified as indicators of good prognosis in the previous analysis, not limited to ER-positive breast cancer. The overexpression of DIABLO appeared to be positive indicator of disease free survival in ER-positive breast cancer only. For further details see, co-pending application Serial No. 60/427090, filed on November 15, 2002, the entire disclosure of which is hereby expressly incorporated by reference.

[0100] The studies described above were performed essentially as described in Example 2 below, except gene amplification was studied using exon-based amplicons. For further details, see copending application Serial No. 60/364,890 filed on March 13, 2002, the entire disclosure of which is hereby expressly incorporated by reference. As attested by the data set forth in Example 2, the data obtained using intron-based amplicons show excellent correlation with the earlier data, and typically provide the added benefit of increased sensitivity.

[0101] The findings of the previous Phase II study of invasive breast ductal carcinoma were subjected to multivariate stepwise analysis, using the Cox Proportional Hazards Model using the following equation:

[0102] RR=exp[coef(geneA) x Ct(geneA) + coef(geneB) x Ct(geneB) + coef(geneC) x Ct(geneC) + ...].

[0103] In this equation, coefficients for genes that are predictors of beneficial outcome are positive numbers and coefficients for genes that are predictors of unfavorable outcome are negative numbers. The "Ct" values in the equation are ΔCts, i.e.

reflect the difference between the average normalized Ct value for a population and the normalized Ct measured for the patient in question. The convention used in the analysis has been that ΔCts below and above the population average have positive signs and negative signs, respectively (reflecting greater or lesser mRNA abundance). The relative risk (RR) calculated by solving this equation indicated if the patient has an enhanced or reduced chance of long-term survival without cancer recurrence.

[0104] In a multivariate analysis, using an interrogation set including a reduced number of genes, the following ten-gene sets have been identified as having particularly strong predictive value of patient survival without cancer recurrence following surgical removal of primary tumor.

1. Bcl2, cyclinG1, NFKBp65, NME1, EPHX1, TOP2B, DR5, TERC, Src, DIABLO;
2. Ki67, XIAP, hENT1, TS, CD9, p27, cyclinG1, pS2, NFKBp65, CYP3A4;
3. GSTM1, XIAP, Ki67, TS, cyclinG1, p27, CYP3A4, pS2, NFKBp65, ErbB3;
4. PR, NME1, XIAP, upa, cyclinG1, Contig51037, TERC, EPHX1, ALDH1A3, CTSL;
5. CA9, NME1, TERC, cyclinG1, EPHX1, DPYD, Src, TOP2B, NFKBp65, VEGFC;
6. TFRC, XIAP, Ki67, TS, cyclinG1, p27, CYP3A4, pS2, ErbB3, NFKBp65.

[0105] In a multivariate analysis, using an interrogation set including all genes identified, the following ten-gene sets have been identified as having particularly strong predictive value of patient survival without cancer recurrence following surgical removal of primary tumor.

1. Bcl2, PRAME, cyclinG1, FOXM1, NFKBp65, TS, XIAP, Ki67, CYP3A4, p27;
2. FOXM1, cyclinG1, XIAP, Contig51037, PRAME, TS, Ki67, PDGFRa, p27, NFKBp65;
3. PRAME, FOXM1, cyclinG1, XIAP, Contig51037, TS, Ki6, PDGFRa, p27, NFKBp65;
4. Ki67, XIAP, PRAME, hENT1, contig51037, TS, CD9, p27, ErbB3, cyclinG1;
5. STK15, XIAP, PRAME, PLAUR, p27, CTSL, CD18, PREP, p53,

RPS6KB1;

6. GSTM1, XIAP, PRAME, p27, Contig51037, ErbB3, GSTp, EREG, ID1, PLAUR;
7. PR, PRAME, NME1, XIAP, PLAUR, cyclinG1, Contig51037, TERC, EPHX1, DR5;
8. CA9, FOXM1, cyclinG1, XIAP, TS, Ki67, NFKBp65, CYP3A4, GSTM3, p27;
9. TFRC, XIAP, PRAME, p27, Contig51037, ErbB3, DPYD, TERC, NME1, VEGFC;
10. CEGP1, PRAME, hENT1, XIAP, Contig51037, ErbB3, DPYD, NFKBp65, ID1, TS.

[0106] Using the same multivariate analysis approach for ER-positive breast cancer, the following ten-gene sets have been identified as having particularly strong predictive value of patient survival without cancer recurrence following surgical removal of primary tumor.

1. PRAME, p27, IGFBP2, HIF1A, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;
2. Contig51037, EPHX1, Ki67, TIMP2, cyclinG1, DPYD, CYP3A4, TP, AIB1, CYP2C8;
3. Bcl2, hENT1, FOXM1, Contig51037, cyclinG1, Contig46653, PTEN, CYP3A4, TIMP2, AREG;
4. HIF1A, PRAME, p27, IGFBP2, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;
5. IGF1R, PRAME, EPHX1, Contig51037, cyclinG1, Bcl2, NME1, PTEN, TBP, TIMP2;
6. FOXM1, Contig51037, VEGFC, TBP, HIF1A, DPYD, RAD51C, DCR3, cyclinG1, BAG1;
7. EPHX1, Contig51037, Ki67, TIMP2, cyclinG1, DPYD, CYP3A4, TP, AIB1, CYP2C8;
8. Ki67, VEGFC, VDR, GSTM3, p27, upa, ITGA7, rhoC, TERC, Pin1;
9. CDC25B, Contig51037, hENT1, Bcl2, HLAG, TERC, NME1, upa, ID1, CYP;
10. VEGFC, Ki67, VDR, GSTM3, p27, upa, ITGA7, rhoC, TERC, Pin1;
11. CTSB, PRAME, p27, IGFBP2, EPHX1, CTSL, BAD, DR5, DCR3, XIAP;

12. DIABLO, Ki67, hENT1, TIMP2, ID1, p27, KRT19, IGFBP2, TS, PDGFB;
13. p27, PRAME, IGFBP2, HIF1A, TIMP2, ILT2, CYP3A4, ID1, EstR1, DIABLO;
14. CDH1; PRAME, VEGFC; HIF1A; DPYD, TIMP2, CYP3A4, EstR1, RBP4, p27;
15. IGFBP3, PRAME, p27, Bcl2, XIAP, EstR1, Ki67, TS, Src, VEGF;
16. GSTM3, PRAME, p27, IGFBP3, XIAP, FGF2, hENT1, PTEN, EstR1, APC;
17. hENT1, Bcl2, FOXM1, Contig51037, CyclinG1, Contig46653, PTEN, CYP3A4, TIMP2, AREG;
18. STK15, VEGFC, PRAME, p27, GCLC, hENT1, ID1, TIMP2, EstR1, MCP1;
19. NME1, PRAM, p27, IGFBP3, XIAP, PTEN, hENT1, Bcl2, CYP3A4, HLAG;
20. VDR, Bcl2, p27, hENT1, p53, PI3KC2A, EIF4E, TFRC, MCM3, ID1;
21. EIF4E, Contig51037, EPHX1, cyclinG1, Bcl2, DR5, TBP, PTEN, NME1, HER2;
22. CCNB1, PRAME, VEGFC, HIF1A, hENT1, GCLC, TIMP2, ID1, p27, upa;
23. ID1, PRAME, DIABLO, hENT1, p27, PDGFRa, NME1, BIN1, BRCA1, TP;
24. FBXO5, PRAME, IGFBP3, p27, GSTM3, hENT1, XIAP, FGF2, TS, PTEN;
25. GUS, HIA1A, VEGFC, GSTM3, DPYD, hENT1, FBXO5, CA9, CYP, KRT18;
26. Bclx, Bcl2, hENT1, Contig51037, HLAG, CD9, ID1, BRCA1, BIN1, HBEGF.

[0107] In view of the excellent correlation between exon-based and intron-based gene expression profiling results (see Example 2), the same gene sets are expected to have similar prognostic value when gene expression profiling is based on the quantitation of RT-PCR signals from intron-based primer/probe sets.

[0108] Further details of the invention will be apparent from the following non-limiting examples.

Example 1

Design and Use of Intron-Specific PCR Primer/Probe Sets

[0109] RNA was extracted from formalin-fixed, paraffin-embedded (FPET) breast cancer biopsy specimens (Clinomics Biosciences Inc., Pittsfield, MA) as follows. Three 10 μM sections were cut and placed in a 1.5 ml tube. Paraffin was removed by xylene extraction (1ml, 3 times) followed by ethanol wash (1ml, twice). RNA was isolated from sectioned tissue blocks using the MasterPure™ Purification kit (Epicentre, Madison, WI). RNA was quantitated by the RiboGreen Fluorescence method (Molecular Probes). Twenty FPET RNA samples were then pooled and used as described below.

[0110] First-strand cDNA was synthesized using Qiagen's Omniscript Reverse Transcriptase with pooled gene specific primers (reverse primers shown in Figure 2) random hexamers and RNase Inhibitor, using pooled FPET RNA (400 ng). A no reverse transcriptase (RT) reaction was also performed with 150 ng of pooled FPET RNA, sufficient RNA to perform the Taqman amplification at 5 ng/well.

Table 2

Reagents	RT Vol (μl)	No RT Vol (μl)	Final conc
10X Buffer RT	4	2	1X
dNTP mix, 5mM each dNTP	4	2	500μM each
ABI Random hexamer, 50μM	1	0.5	1.25μM
GSP pool,1μM	2	1	50nM
ABI RNase Inhibitor, 20U/μl	1	1	20U/rxn
Omniscript RT, 4U/μl	2	0	8U or 0U/rxn
Nuclease free water	10	5.5	
Pooled FPET RNA (164 ng/μl)	16	8	65.6 ng/μl
Total vol	40	20	
Reaction conditions:	37 °C, 60 min, 93°C, 5 min		

TaqMan Assay

[0111] TaqMan assays for the 48-gene panel were carried out in triplicate wells with reaction volume of 25 μl and RNA input of 5 ng per assay. A "no RT "

reaction for each gene was carried out in a single well as a control to verify that RNA rather than DNA signals were being measured. Real time quantitation was performed on the ABI 7700 using the following parameters:

[0112] Cycling conditions: 95°C, 10 min for one cycle, 95°C, 20 sec followed by 60°C, 45 sec, 40 cycles.

[0113] Volume reaction: 25 μl.

[0114] Dye layer setting: FAM, (the passive reference is ROX)

Results

[0115] Intron specific Taqman primer-probe sets were designed based on masked introns of CEGP1, FOXM1, PRAME and STK15.genes, To delineate intron sequences within the genes, the NCBI reference sequence for each mRNA (NM_XXXXXX) was aligned to the human genome using the BLAST-like alignment tool (BLAT) program available at the University of Santa Cruz on-line genome resource site (http://genome.ucsc.edu). Intron sequences were then searched for repetitive sequences using the Repeat Masker program available on-line through the Baylor College of Medicine (http://searchlauncher.bcm.tmc.edu/seq-util/seq-util.html). Repeat sequences, such as Alu repeats, are identified by this program and masked. It is important to exclude these sequences prior to designing primer-probes because they yield strong, non-specific signals. The masked intron sequences (Figures 1A-M) were then used to design Taqman primer-probe sets using Primer Express (ABI). Other programs suitable for primer-probe sets include, for example, the newer primer probe design program for MGB assays-by-design (ABI). The amplicons for each primer-probe set are delineated in bold font in Figure 1. Each specific primer-probe set is shown in Figure 2.

[0116] The intron-specific primer-probe sets (test genes) were used together with their corresponding exon-specific primer-probe set (references gene) in standard Taqman gene expression profile experiments using pooled FPET RNA. Normalized expression was calculated by the formula $2^{\Delta Ct}$ where ΔCt is the difference between the Cts of the test gene primer-probe set and the reference gene primer-probe sets [Ct (reference)-Ct (test)].

Example 2

A Phase II Study of Gene Expression in Premalignant and Malignant Breast Tumors

[0117] A gene expression study was designed and conducted with the primary goal to molecularly characterize gene expression in paraffin-embedded, fixed tissue samples of invasive breast ductal carcinoma, and to explore the correlation between such molecular profiles and disease-free survival.

Study design

[0118] Molecular assays were performed on paraffin-embedded, formalin-fixed primary breast tumor tissues obtained from 60 individual patients diagnosed with breast cancer. All patients underwent surgery with diagnosis of invasive carcinoma of the breast. Patients were included in the study only if histopathologic assessment, performed as described in the Materials and Methods section, indicated adequate amounts of tumor tissue and homogeneous pathology.

Materials and Methods

[0119] Each representative tumor block was characterized by standard histopathology for diagnosis, semi-quantitative assessment of amount of tumor, and tumor grade. A total of 6 sections (10 microns in thickness each) were prepared and placed in two Costar Brand Microcentrifuge Tubes (Polypropylene, 1.7 mL tubes, clear; 3 sections in each tube). If the tumor constituted less than 30% of the total specimen area, the sample may have been crudely dissected by the pathologist, using gross microdissection, putting the tumor tissue directly into the Costar tube.

[0120] If more than one tumor block was obtained as part of the surgical procedure, all tumor blocks were subjected to the same characterization, as described above, and the block most representative of the pathology was used for analysis.

Gene Expression Analysis

[0121] mRNA was extracted and purified from fixed, paraffin-embedded tissue samples, and prepared for gene expression analysis as described above.

[0122] Molecular assays of quantitative gene expression were performed by RT-PCR, using the ABI PRISM 7900™ Sequence Detection System™ (Perkin-Elmer-Applied Biosystems, Foster City, CA, USA). ABI PRISM 7900™ consists of a thermocycler, laser, charge-coupled device (CCD), camera and computer. The system

amplifies samples in a 384-well format on a thermocycler. During amplification, laser-induced fluorescent signal is collected in real-time through fiber optics cables for all 384 wells, and detected at the CCD. The system includes software for running the instrument and for analyzing the data.

Analysis and Results

[0123] Tumor tissue was analyzed for expression of 48 different RNA sequences representing products of 37 different genes. The threshold cycle (Ct) values for each patient were normalized based on the median of all genes for that particular patient. Clinical outcome data were available for all patients from a review of registry data and selected patient charts.

[0124] Outcomes were classified as:

0 died due to breast cancer or to unknown cause or alive with breast cancer recurrence;

1 alive without breast cancer recurrence or died due to a cause other than breast cancer

[0125] Analysis was performed by:

[0126] Analysis of the relationship between normalized gene expression and the time to outcome (0 or 1 as defined above) where patients who were alive without breast cancer recurrence or who died due to a cause other than breast cancer were censored. This approach was used to evaluate the prognostic impact of individual genes and also sets of multiple genes.

[0127] For each gene a Cox Proportional Hazards model (see, e.g. Cox, D. R., and Oakes, D. (1984), *Analysis of Survival Data*, Chapman and Hall, London, New York) was defined with time to recurrence or death as the dependent variable, and the expression level of the gene as the independent variable. The genes that have a p-value < 0.05 in the Cox model were identified. For each gene, the Cox model provides the relative risk (RR) of recurrence or death for a unit change in the expression of the gene. One can choose to partition the patients into subgroups at any threshold value of the measured expression (on the Ct scale), where all patients with expression values above the threshold have higher risk, and all patients with expression values below the threshold have lower risk, or vice versa, depending on whether the gene is an indicator of poor (RR>1.01) or good (RR<1.01) prognosis. Thus, any threshold value will define subgroups of patients with respectively increased or decreased risk.

[0128] Table 3, below, shows pairwise correlation of expression (presented by correlation coefficients) between the tested introns and exons for the genes CEGP1, FOXM1, PRAME, and STK15. For two of the four genes, CEGP1 and PRAME, introns were found that yielded correlation coefficients [for co-expression with their respective exons] above 0.90. In the case of STK15, one intron correlated with exon expression with a correlation coefficient ~0.80. For FOXM1, intron:exon expression correlations were significantly lower. In this last case, however, it seems likely that actual expression may be highly correlated but not detectable for a technical reason. Expression of the FOXM1 exon in many patients was beneath the detection threshold of the assay, which potentially prevents detection of high correlations that may exist. If this hypothesis is correct, FOXM1 introns would still register as negative clinical prognostic markers as previously demonstrated for FOXM1. As shown later, this result occurs.

[0129] Figures 3, 4 and 5 show that the pairwise correlation of expression of the tested RNAs against CEGP1, PRAME and STK15 exon RNAs. As shown, respective introns of these genes yielded the highest correlations. It is noteworthy that the panel of 48 genes included genes that we selected, by several bioinformatics-based strategies, as particularly likely to correlate in expression with CEGP1, PRAME, STK15, and FOXM1. Those non-intron-based strategies were most successful in the case of STK15, as several candidate genes had expression correlation coefficients in the range of 0.6-0.7.

Table 3. Correlations between Intron and Exon Expression for Four Genes

	ExpressionCorrelation Coefficient {R}				
	CEGP1intron1.1	*CEGP1intron3.1*	*CEGP1intron4.1*	*CEGP1intron5.1*	*CEGP1.2*
CEGP1intron1.1	1.00				
CEGP1intron3.1	0.89	1.00			
CEGP1intron4.1	0.97	0.82	1.00		
CEGP1intron5.1	0.91	0.87	0.88	1.00	
CEGP1.2	0.91	0.80	0.90	0.87	1.00

	FOXM1intron3.3	FOXM1intron5.1	FOXM1intron7.1	FOXM1.1
FOXM1intron3.3	1.00			
FOXM1intron5.1	0.48	1.00		
FOXM1intron7.1	0.54	0.73	1.00	
FOXM1.1	0.44	0.33	0.38	1.00

	STK15intron1.1	STK15intron2.1	STK15intron4.1	STK15.2
STK15intron1.1	1.00			
STK15intron2.1	0.78	1.00		
STK15intron4.1	0.69	0.74	1.00	
STK15.2	0.63	0.70	0.78	1.00

	PRAMEintron2.1	PRAME.3
PRAMEintron2.1	1.00	
PRAME.3	0.97	1.00

[0130] Table 4, below, shows the impact upon patient survival of expression of CEGP1, FOXM1, PRAME, and STK15, exons and introns. The parent exons all had statistically significant impact on relative risk [RR], as we previously determined, except in the case of FOXM1. Because the present study evaluated 60 patients from the original 146 patient group, the FOXM1 marker may have fallen from significance because the statistical hazard of examining a reduced data set. Very notably, for all four tested genes, intron expression significantly impacted RR, and in the same direction as the parent exons.

Table 4. Cox Model Results for 60 Patients with Breast Cancer

		Prognostic Correlations			
Gene	Coef	RR=exp(coef)	se(coef)	z	p
CEGP1.2	-0.202	0.817	0.050	-4.024	0.00006
CEGP1intron1.1	-0.329	0.720	0.087	-3.771	0.00016
CEGP1intron3.1	-0.261	0.770	0.078	-3.335	0.00085
CEGP1intron4.1	-0.275	0.760	0.073	-3.774	0.00016
CEGP1intron5.1	-0.312	0.732	0.082	-3.817	0.00014
FOXM1.1	0.175	1.192	0.136	1.289	0.19700
FOXM1intron3.3	0.304	1.355	0.120	2.523	0.01160
FOXM1intron5.1	0.514	1.673	0.195	2.639	0.00832
FOXM1intron7.1	0.546	1.726	0.182	2.993	0.00276
PRAME.3	0.125	1.133	0.054	2.294	0.02180
PRAMEintron2.1	0.125	1.133	0.052	2.397	0.01650
STK15.2	0.692	1.998	0.201	3.450	0.00056
STK15intron1.1	0.357	1.429	0.149	2.400	0.01640
STK15intron2.1	0.391	1.479	0.154	2.536	0.01120
STK15intron4.1	0.410	1.506	0.133	3.084	0.00204

[0131] A common perception exists that steady state levels of transcribed exon sequences greatly exceed those of transcribed intron sequences (Sharp *et al. Ann. Rev. Biochem.* 55: 1119-50 [1986]). Nevertheless, our examination of CEGP1, FoxM1, PRAME and STK15 exon and intron expression, using TaqMan[TM] RT-PCR to assay RNA from fixed paraffin-embedded breast cancer tissue, demonstrated that intron and exon signal intensities were in the same range, and in all cases in the useful detection range of the assay [data not shown]. The detection of intronic RNA in this study is all the more notable because the tissue used was fixed in formalin, which degrades RNA, and thus substantially limits the ability to detect RNA (T.E. Godfrey et al. J. Mol. Diag. 2: 84-91 [2000]). In the case of CEGP1 three of the tested introns yielded lower signals and one a higher signal than the exon. In the case of FOXM1, five of nine tested introns yielded higher signals than the exon. In the case of PRAME signal intensities from the tested intron and exon were nearly identical. Finally, for STK15 all introns had signal intensities that were 1/4 to 1/20 those of the exon, but were still in the useful range of the assay. Thus, these results indicate that steady state levels of expressed introns are adequate for use of intron RNAs as molecular markers.

[0132] All references cited throughout the disclosure, including the examples, are hereby expressly incorporated by reference for their entire disclosure.

[0133] While the present invention has been described with reference to what is considered to be specific embodiments, it is to be understood that the invention is not so limited. To the contrary, the invention is intended to cover various modifications and equivalents included within the spirit and scope of the appended claims. For example, while the disclosure includes various breast cancer-associated genes and gene sets, similar genes and gene sets and methods concerning other types of cancer are specifically within the scope herein.